

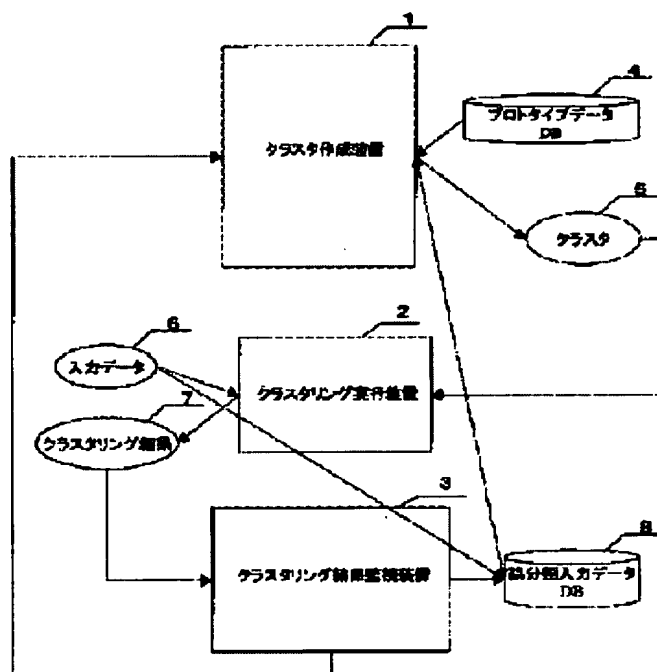
CLUSTERING DEVICE

Patent number: JP2001283184
Publication date: 2001-10-12
Inventor: NAKAMITSU HIROAKI
Applicant: MATSUSHITA ELECTRIC IND CO LTD
Classification:
 - International: G06N3/00; G06F9/44; G06F17/30
 - european:
Application number: JP20000091863 20000329
Priority number(s):

Abstract of JP2001283184

PROBLEM TO BE SOLVED: To provide a clustering device capable of coping with the dynamic change of data in clustering in a simple constitution and procedure.

SOLUTION: This clustering device for classifying input data by using a cluster is provided with a cluster preparing device 1 for preparing a cluster, a clustering performing device 2 for performing the clustering of the input data by using the cluster prepared by the cluster preparing device, a clustering result monitoring device 3 for monitoring the clustering result of the clustering performing device, and for identifying the erroneously classified input data, and a storage means 8 for storing the erroneously classified input data. When the fixed number of data or more are stored in the storage means, a new cluster is prepared by the cluster preparing device based on the data. Thus, it is possible to correct the cluster corresponding to the dynamic change of the input data, and to reduce the erroneous classification.



Data supplied from the esp@cenet database - Patent Abstracts of Japan

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号
特開2001-283184
(P2001-283184A)

(43) 公開日 平成13年10月12日 (2001. 10. 12)

(51) Int.Cl. ⁷	識別記号	F I	テマコード (参考)
G 0 6 N 3/00	5 6 0	G 0 6 N 3/00	5 6 0 A 5 B 0 7 5
G 0 6 F 9/44	5 8 0	G 0 6 F 9/44	5 8 0 A
17/30	2 1 0	17/30	2 1 0 D

審査請求 未請求 請求項の数 5 O L (全 7 頁)

(21) 出願番号 特願2000-91863(P2000-91863)

(22) 出願日 平成12年3月29日 (2000. 3. 29)

(71) 出願人 000005821

松下電器産業株式会社

大阪府門真市大字門真1006番地

(72) 発明者 仲光 廣晃

大阪府門真市大字門真1006番地 松下電器
産業株式会社内

(74) 代理人 100099254

弁理士 役 昌明 (外 3 名)

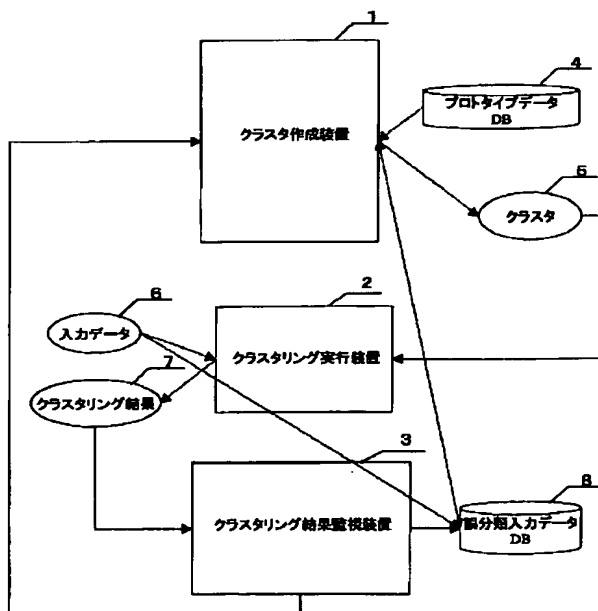
Fターム (参考) 5B075 NR12

(54) 【発明の名称】 クラスタリング装置

(57) 【要約】

【課題】 簡単な構成と手順で、クラスタリングにおけるデータの動的変化に対応することができるクラスタリング装置を提供する。

【解決手段】 入力データを、クラスタを用いて分類するクラスタリング装置において、クラスタを作成するクラスタ作成装置1と、クラスタ作成装置により作成されたクラスタを用いて、入力データのクラスタリングを実行するクラスタリング実行装置2と、クラスタリング実行装置のクラスタリング結果を監視して誤分類された入力データを識別するクラスタリング結果監視装置3と、誤分類された入力データを蓄積する蓄積手段8とを設け、蓄積手段に一定数以上のデータが蓄積された場合に、このデータを基に、クラスタ作成装置が新たなクラスタを作成するように構成している。入力データの動的変化に対応してクラスタを修正し、誤分類を抑えることができる。



【特許請求の範囲】

【請求項1】 入力データを、クラスタを用いて分類するクラスタリング装置であって、
前記クラスタを作成するクラスタ作成装置と、
前記クラスタ作成装置により作成されたクラスタを用いて、入力データのクラスタリングを実行するクラスタリング実行装置と、
前記クラスタリング実行装置のクラスタリング結果を監視して誤分類された入力データを識別するクラスタリング結果監視装置と、
誤分類された前記入力データを蓄積する蓄積手段とを備え、
前記蓄積手段に一定数以上のデータが蓄積された場合に、前記データを基に、前記クラスタ作成装置が新たなクラスタを作成することを特徴とするクラスタリング装置。

【請求項2】 前記クラスタ作成装置は、前記蓄積手段に一定数以上のデータが蓄積された場合に、前記データを用いて新たなクラスタを自動的に作成し、既に作成したクラスタに追加することを特徴とする請求項1に記載のクラスタリング装置。

【請求項3】 前記クラスタリング結果に、クラスタリングの誤差のデータが含まれることを特徴とする請求項1に記載のクラスタリング装置。

【請求項4】 前記クラスタ作成装置は、プロトタイプデータを入力として自己組織化マップを生成する自己組織化マップ生成手段と、生成された前記自己組織化マップを区分しクラスタを形成するクラスタ形成手段とを備えることを特徴とする請求項1に記載のクラスタリング装置。

【請求項5】 前記クラスタリング結果監視装置は、クラスタリング結果を監視するクラスタリング結果監視手段と、前記蓄積手段に一定数以上のデータが蓄積された場合に、前記データを入力として自己組織化マップを生成する自己組織化マップ修正手段とを備え、前記クラスタ作成装置のクラスタ形成手段は、前記自己組織化マップ修正手段が自己組織化マップを生成した場合、その自己組織化マップを区分してクラスタを作成し、既に作成したクラスタに追加することを特徴とする請求項4に記載のクラスタリング装置。

【発明の詳細な説明】**【0001】**

【発明の属する技術分野】 本発明は、多数のデータをその類似性からクラスに分類するクラスタリング装置に関し、特に、入力データの動的変化に適切に対応できるようにしたものである。

【0002】

【従来の技術】 従来、クラスタリング手法として、さまざまなものが提案されている。図6には、最も一般的なクラスタリング装置の例を示している。

【0003】 ここで、100は学習のためのプロトタイプデータ群を示し、102、103は、プロトタイプデータ群の個々のデータを初期クラスタとみなしたクラスタA、クラスタBを示し、104は、クラスタA102とクラスタB103との距離を示し、105はクラスタA102とクラスタB103とを統合したクラスタCを示す。200は、プロトタイプデータから作成されたクラスタ結果を示し、201、202は最終的に作成されたクラスタY、クラスタZを示す。300は、クラスタを用いたクラスタリング装置を示し、301は201とまったく同型のクラスタY、302は202とまったく同型のクラスタZを示し、303は、クラスタリングの対象である入力Xを示し、304はクラスタが存在する空間上の入力X303のポイントを示す。

【0004】 この装置では、まず、クラスタリング装置に必要なクラスタを作成する。これは以下の作業により求められる。

【0005】 学習のためのプロトタイプデータ群100から、最も距離の近いクラスタを探し、その結果、クラスタA102とクラスタB103とが選ばれたとすると、この2つを統合してクラスタC105とし、クラスタA、Bは削除する。この時クラスタC105は、クラスタA102とクラスタB103との値を両方ともを持つ。次に、また同様にプロトタイプデータ群100から、最も距離の近いクラスタを探し、それらを統合する、という一連の作業を繰り返す。この時、全クラスタ数が1になった場合や、最も距離の近いクラスタ同士の距離が、ある一定値より大きかった場合は、作業を終了する。

【0006】 この一連の作業により、プロトタイプデータから作成されたクラスタ結果200が求められ、最終的に統合されたクラスタがクラスタY201、クラスタZ202となる。

【0007】 これら最終的に統合されたクラスタを用い、実際のクラスタリングを行うのがクラスタを用いたクラスタリング装置300である。このクラスタを用いたクラスタリング装置300に入力X303が入力された時、入力X303がクラスタY301内に含まれる時、入力X303は、クラスタY301にクラスタリングされたという結果となる。

【0008】 また、クラスタリングに自己組織化マップ（SOM：Self-Organization Map、詳しくは、T. Kohonen, "Self-Organization and Associative Memory", Third Edition, Springer-Verlag, Berlin, 1989に記載されている。）と呼ばれるニューラルネットワークを用いる手法も知られている（特開平7-234853号）。この方法では、プロトタイプデータをSOMに入力して、SOMを形成するニューロンを学習し、学習したニューロンをクラスタに分類する。クラスタが形成された後、SOMに入力データを与えると、その入力に近い値を持つニューロンが決定され、入力データがクラスタリングされる。

【0009】

【発明が解決しようとする課題】しかし、前述のようなクラスタリング手法では、プロトタイプデータを用いてクラスタを形成しているため、プロトタイプデータにのみ偏ったクラスタが形成される。そのため、実際にこれらのクラスタを用いて実データのクラスタリングを行った時、入力データの動的な変化に対応できない、と云う問題点がある。

【0010】つまり、新たなクラスに属すべきデータが、時間の経過とともに生じた場合などに、従来の方法では、全く対応ができず、いずれかのクラスタに誤分類されることになる。

【0011】この誤分類を防ぐためには、従来の方式では、プロトタイプデータも含めて、すべてのデータを用いてクラスタリングし直す必要があり、大きな作業負担が強えられる。データを新たに追加した場合にクラスタの修正を行う方法が、特開平5-205058号に開示されているが、これは、新たなデータを追加したことが既知でなければならず、かつ外部からデータの追加によるクラスタの修正を実行することを知らせる必要があり、追加するデータを自動的に集めたり、クラスタを自動的に修正することはできない。

【0012】本発明は、こうした従来の問題点を解決するものであり、簡単な構成と手順で、クラスタリングにおけるデータの動的変化に対応することができるクラスタリング装置を提供することを目的としている。

【0013】

【課題を解決するための手段】そこで、本発明では、入力データを、クラスタを用いて分類するクラスタリング装置において、クラスタを作成するクラスタ作成装置と、クラスタ作成装置により作成されたクラスタを用いて、入力データのクラスタリングを実行するクラスタリング実行装置と、クラスタリング実行装置のクラスタリング結果を監視して誤分類された入力データを識別するクラスタリング結果監視装置と、誤分類された入力データを蓄積する蓄積手段とを設け、蓄積手段に一定数以上のデータが蓄積された場合に、このデータを基に、クラスタ作成装置が新たなクラスタを作成するように構成している。

【0014】そのため、入力データの動的変化に対応してクラスタを修正し、誤分類を抑えることができる。

【0015】

【発明の実施の形態】以下、本発明の実施の形態について、図面を用いて説明する。なお、本発明はこれら実施の形態に何等限定されるものではなく、その要旨を逸脱しない範囲において種々なる態様で実施し得る。

【0016】（第1の実施形態）第1の実施形態のクラスタリング装置は、図1に示すように、プロトタイプデータを管理するプロトタイプデータDB4と、プロトタイプデータを用いてクラスタ5を作成するクラスタ作成

装置1と、作成されたクラスタ5を用いて入力データ6をクラスタリングするクラスタリング実行装置2と、クラスタリング実行装置2のクラスタリング結果7を監視するクラスタリング結果監視装置3と、クラスタリング結果監視装置3によって誤分類と判断されたデータを管理する誤分類入力データDB8とを備えている。

【0017】この装置では、クラスタ作成装置1が、プロトタイプデータDB4を用いてクラスタ5を生成する。クラスタリング実行装置2は、生成されたクラスタ5を用いて、入力された入力データ6をクラスタリングし、クラスタリング結果7を出力する。クラスタリング結果監視装置3は、出力されたクラスタリング結果7を監視し、入力データ6のクラスタリング結果7に含まれる誤差が、ある一定値以上の値であり、明らかに誤分類であると判断した時、その入力データ6を誤分類入力データDB8に追加し、誤分類入力データDB8に溜まったデータの数をカウントする。この誤分類入力データDB8内のデータがある一定数を超えた時、クラスタ作成装置1に、この誤分類入力データDB8を用いて、クラスタを作成するように指示する。

【0018】各装置の動作をさらに詳しく説明する。まず、クラスタ作成装置1は、クラスタ5が作成されていない時と、クラスタリング結果監視装置3からクラスタの作成を指示された時に動作する。

【0019】クラスタ5が作成されていない時は、プロトタイプデータDB4内のプロトタイプデータ群の個々のデータを初期クラスタと見なし、その中から、最も距離の近いクラスタを探す。この距離は図5の式1によって求める。この時求められた2つのクラスタを統合し新たなクラスタとする。統合されたクラスタは削除し、また新たに作られたクラスタは、統合により削除されたクラスタの値をすべて持つ。同様にまたプロトタイプデータDB4から、最も距離の近いクラスタを探し、それらを統合する、という一連の作業を繰り返す。この時、全クラスタ数が1になった場合や、最も距離の近いクラスタ同士の距離が、ある一定値より大きかった場合は、作業を終了する。

【0020】この一連の作業により、プロトタイプデータ4から作成されたクラスタ5を作成する。

【0021】次に、クラスタリング結果監視装置3からクラスタ作成の指示を受けた時は、誤分類入力データDB8を用い、クラスタ5を作成するのと同じ動作で、クラスタを作成する。この時、作成されたクラスタで、クラスタ内に含まれる値の数が一定以上のものを新たなクラスタとしてクラスタ5に加える。最後に、誤分類入力データDB8をクリアする。

【0022】次に、クラスタリング実行装置2の動作について説明する。クラスタ作成装置1により作成されたクラスタ5を用いて、入力された入力データ6と、距離の最も近いクラスタを選択する。この距離の計算は、図

5の式1によって求める。この時、選択されたクラスタと、誤差を表す、計算された距離とをクラスタリング結果7として出力する。

【0023】クラスタリング結果監視装置3は、出力されたクラスタリング結果7に含まれる誤差、即ち、計算された距離が、ある一定値以上の値である時、誤分類入力データDB8に入力データ6を追加しその数をカウントし、この誤分類入力データDB8内のデータがある一定数を超えた時、クラスタ作成装置1にこの誤分類入力データDB8を用いて、クラスタを作成するように指示する。

【0024】以上のように、この実施形態のクラスタリング装置では、稼動中にもクラスタの自動作成が可能であり、入力データの動的な変化に対応して自動的にクラスタを作成することができる。そのため、入力データの動的な変化に起因する誤分類の発生が迅速に抑えられる。また、この装置では、クラスタの再作成が、実データのクラスタリングの過程で誤分類データとして自動収集されたデータのみを用いて行われるため、少ない負担でクラスタの修正を実行することができる。

【0025】(第2の実施形態)第2の実施形態のクラスタリング装置は、自己組織化マップ(以下、SOMと云う)を利用してクラスタを作成する。

【0026】この装置は、図2に示すように、第1の実施形態と同様、プロトタイプデータDB4、クラスタ作成装置1、クラスタリング実行装置2、クラスタリング結果監視装置3及び誤分類入力データDB8から成り、クラスタ作成装置1は、プロトタイプデータを入力するデータ入力手段11と、SOM9を作成するSOM作成手段12と、SOM9を用いてクラスタを生成するクラスタ生成手段13とを備え、また、クラスタリング結果監視装置3は、クラスタリング実行装置2のクラスタリング結果7を監視するクラスタリング結果監視手段31と、誤分類入力データDB8のデータを用いてSOM10を作成するSOM修正手段32とを備えている。

【0027】この装置では、クラスタ作成装置1のデータ入力手段11がプロトタイプデータDB4からデータを入力し、このデータを用いてSOM作成手段12がSOM9を作成し、クラスタ生成手段13が、SOM9を用いてクラスタ5を生成する。クラスタリング実行装置2は、生成されたクラスタ5を用いて入力された入力データ6をクラスタリングし、クラスタリング結果7を出力する。クラスタリング結果監視装置3のクラスタリング結果監視手段31は、出力されたクラスタリング結果7に含まれる誤差が、ある一定値以上の値であり、明らかに誤分類であると判断した時、誤分類入力データDB8に入力データ6を追加し、その数をカウントする。

【0028】誤分類入力データDB8内のデータがある一定数を超えた時、SOM修正手段32は、誤分類入力データDB8のデータを入力として新たなSOM10を作成

し、クラスタ作成手段13にSOM10を用いたクラスタ作成を指示する。これを受けて、クラスタ作成手段13は、SOM10を用いてクラスタを作成し、既に作成されているクラスタ5に追加する。

【0029】次に、各部の動作についてさらに詳しく説明する。まず、SOM作成手段12の動作について説明する。

【0030】SOMは、図4に示すように、2次元上に配置されたニューロン402から形成され、各ニューロンは、参照ベクトル403と呼ばれる入力と同じ次元のベクトルを持つ。

【0031】SOM作成手段12は、図3のフローチャートに示す手順でSOMを作成する。

ステップA1: 学習回数Tを0にセットし、

ステップA2: 図4のように2次元上に配置したニューロンを作成し、各ニューロンに対し、入力と同じ次元の参照ベクトルを乱数で与える。

【0032】ステップA3: データ入力手段11がプロトタイプデータDB4からランダムでデータの一つを取り出す。

【0033】ステップA4: このデータに対して、図5の式(2)を満たす参照ベクトルを持つニューロンCを決定する。

【0034】ステップA5: ニューロンCの近傍に位置するニューロンの参照ベクトルを、図5の式(3)に従って更新する。

【0035】ステップA6: 学習回数Tが規定した回数に達した場合には、

ステップA8: 終了する。

【0036】ステップA6において、学習回数Tが規定回数に達していない場合には、

ステップA7: 学習回数Tの値を一つ増やし、ステップA2に戻る。

【0037】次に、クラスタ生成手段13は、クラスタ5が作成されていない時と、SOM修正手段32からクラスタの作成の指示を受けた時に動作する。

【0038】まず、クラスタ5が作成されていない時、SOM9を用いてクラスタ5を作成する。SOM9の各ニューロンに対し、図5の式(4)を満たす参照ベクトルを持つニューロンを選択し、選択されたニューロンを初期クラスタと見なす。その中から、最も距離の近いクラスタを探す。この距離は図5の式(1)によって求める。この時求められた2つのクラスタを統合し新たなクラスタとする。統合されたクラスタは削除し、また新たに作られたクラスタは、統合により削除されたクラスタの値をすべて持つ。同様にまた、最も距離の近いクラスタを探し、それらを統合する、という一連の作業を繰り返す。この時、全クラスタ数が1になった場合や、最も距離の近いクラスタ同士の距離が、ある一定値より大きかった場合は、作業を終了する。

【0039】また、SOM修正手段32からクラスタの作成の伝達を受けた時も同様に、SOM10を用いてクラスタを作成し、クラスタ5に追加をする。

【0040】クラスタリング実行装置2は、第1の実施形態と同様、クラスタ5を用いて入力データ6をクラスタリングし、クラスタリング結果7を出力する。クラスタリング結果監視手段31は、出力されたクラスタリング結果7に含まれる誤差が、ある一定値以上の値であり、明らかに誤分類であると判断した時、誤分類入力データDB8に入力データ6を追加し、その数をカウントする。

【0041】この誤分類入力データDB8内のデータがある一定数を超えた時、SOM修正手段32は、誤分類入力データDB8のデータを入力として、図3のフローチャートに従って、マップの大きさがSOM9の縦または横のニューロンの数と等しい、小さいSOM10を作成する。そして、誤分類入力データDB8をクリアし、クラスタ作成手段13にクラスタの作成を指示する。クラスタ作成手段13は、前述するように、SOM10を用いてクラスタを作成し、作成済みのクラスタ5に追加する。

【0042】以上のように、この実施形態のクラスタリング装置では、SOMを用いてクラスタリングを行っているため、既存のSOMをそのまま適用することができ、さらに新たにクラスタを作成する際に非常に小さいSOMを用いるので処理速度も高く、その実用的効果は大きい。また、この新たなクラスタの作成には、実データのクラスタリングの過程で誤分類データとして自動収集されたものが使用されるため、この新たなクラスタの作成により、入力データの動的な変化に対応することができる。

【0043】

【発明の効果】以上の説明から明らかなように、本発明のクラスタリング装置は、入力データの動的な変化に対応して、新たなクラスタを速やかに作成することができ、入力データの動的な変化に起因する誤分類の発生を抑えることが可能である。

【0044】また、この新たなクラスタの作成は、クラスタリングを実行したときに、誤分類データとして自動収集されたデータだけを用いて行われるため、その作成負担は少なく済む。

【0045】また、誤分類されたデータからクラスタを直接作成する手段を持つ装置では、装置稼動中にもクラスタを自動で作成することが可能であり、入力データの動的な変化に素早く対応できるという有利な効果が得られる。

【0046】また、SOMを用いてクラスタリングする装置では、既存のSOMをそのまま適用することができ、また、新たにクラスタを作成する際には非常に小さいSOMを用いるので処理速度も高いという有効な効果が得られる。

【0047】このことにより、本発明は、入力データが時間的に変化するものをクラスタリングする装置に適用して効果を発揮することができ、例えば、時間的に変化する生徒の学習結果を入力データとして生徒を分類する学習システムのクラスタリング装置や、インターネットのホームページにアクセスする視聴者の嗜好性を調査するクラスタリング装置などに用いた場合に、極めて有効である。

【図面の簡単な説明】

【図1】本発明の第1の実施形態におけるクラスタリング装置の構成を表すブロック図、

【図2】本発明の第2の実施形態におけるクラスタリング装置の構成を示すブロック図、

【図3】第2の実施形態においてSOM作成の手順を示すフローチャート、

【図4】SOMを視覚的に示す図、

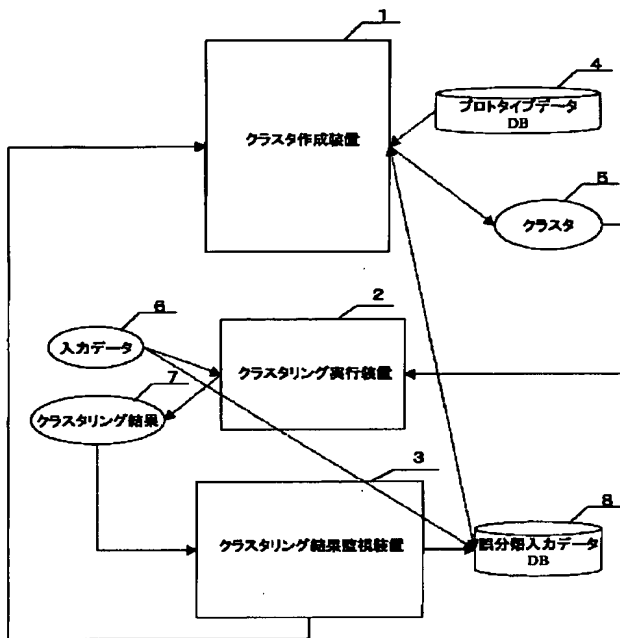
【図5】数式を示す図、

【図6】従来のクラスタリング装置の一例を示す図である。

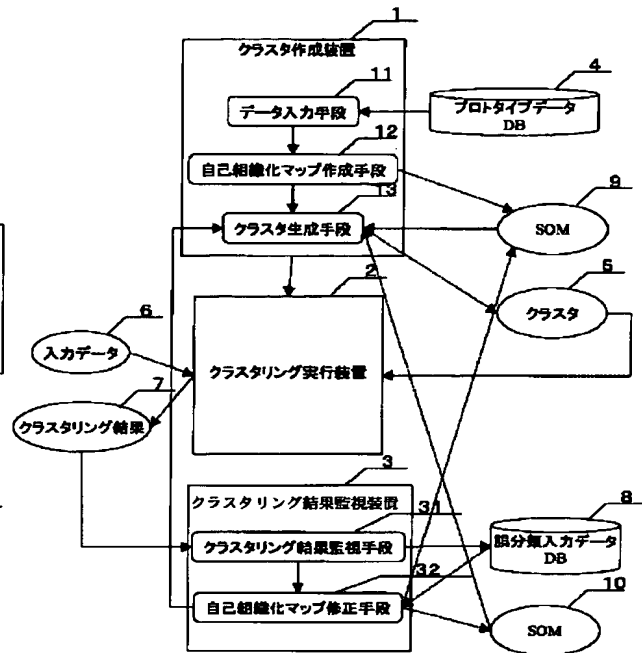
【符号の説明】

- 1 クラスタ作成装置
- 2 クラスタリング実行装置
- 3 クラスタリング結果監視装置
- 4 プロトタイプデータDB
- 5 クラスタ
- 6 入力データ
- 7 クラスタリング結果
- 8 誤分類入力データDB
- 9、10 SOM
- 11 データ入力手段
- 12 SOM作成手段
- 13 クラスタ生成手段
- 31 クラスタリング結果監視手段
- 32 SOM修正手段
- 100 プロトタイプデータ群
- 102 クラスタA
- 103 クラスタB
- 104 距離
- 105 クラスタC
- 200 プロトタイプデータから作成されたクラスタ結果
- 201 クラスタY
- 202 クラスタZ
- 300 クラスタを用いたクラスタリング装置
- 301 クラスタY
- 302 クラスタZ
- 303 入力X
- 304 入力Xのポイント
- 401 SOM
- 402 ニューロン
- 403 参照ベクトル

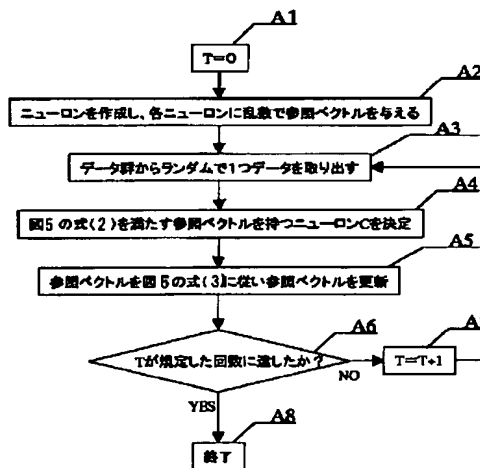
【図1】



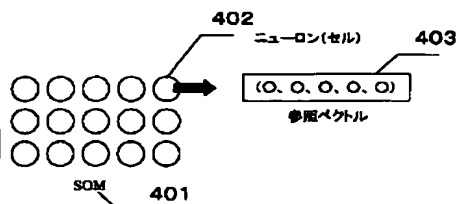
【図2】



【図3】



【図4】



【図5】

$$N\text{次元のデータ}X,Y\text{の距離}D \quad D = \sqrt{\sum_{l=0}^N (x_l - y_l)^2} \quad (1)$$

$$C = \arg \min_i \|x - m_i\| \quad (2)$$

i :ニューロン番号

X :入力データ

m_i :ニューロン i の参照ベクトル

$\|\cdot\|$:式(1)によって求められる距離

$$m_i = \begin{cases} m_i + h_{ci}(t)[x(t) - m_i(t)] & i \in N_c \\ m_j & i \notin N_c \end{cases} \quad (3)$$

$$h_{ci} = \alpha(t) \cdot \exp \left(-\frac{\|x_c - r_i\|^2}{2\sigma^2(t)} \right)$$

N_c :ニューロン C の近傍に位置するニューロン

i :ニューロン番号

α :学習係数: $0 < \alpha < 1$ (時間とともに減少)

σ :近傍幅(時間とともに減少)

$\|\cdot\|$:式(1)によって求められる距離

r_c :2次元マップ上のニューロン C の位置ベクトル

r_i :2次元マップ上のニューロン i の位置ベクトル

$$d_i < \min_{j \in N_i} d_j \quad (4)$$

$$d_i = \frac{1}{|N_i|} \sum_{j \in N_i} \|m_i - m_j\|$$

N_i :ニューロン i の4近傍中に存在するニューロンの集合

m_i :ニューロン i の参照ベクトル

m_j :ニューロン j の参照ベクトル

$\|\cdot\|$:式(1)によって求められる距離

【図6】

